CLAIMS

What is claimed is:

1. A method for generating clusters with user perspective comprising:
receiving session logs;
performing log-based clustering on the session logs to generate session clusters;
representing each session cluster as a log-based document suitable for content based clustering;
receiving a plurality of documents that includes a first document that was accessed in one session and a second document that was not accessed in the sessions;
replacing the first document with a log-based document associated with the session cluster that includes the first document; and
performing content based clustering on at least the first document and the second document to generate clusters with user perspective.

2. A method for generating clusters of claim 1 wherein
representing each session cluster as a log-based document suitable for content based clustering includes modifying each document referenced in the session cluster so that the Euclidean distance between the documents is the same.

3. A method for clustering documents comprising the steps of:
generating a hybrid matrix of vectors comprising a first vector representing a first document and a second vector representing a log-based document cluster; and
clustering the documents using the hybrid matrix.

4. The method recited in claim 3 wherein the step of generating the hybrid matrix comprises the steps of:
accessing retrieval session logs;

clustering retrieval sessions into session clust rs;
generating, for each session cluster, a log-based document cluster by combining all documents opened during any retrieval session of the session cluster;
generating a log-based document cluster vector for each of the log-based document clusters;
replacing each document in the log-based document cluster with the log-based document cluster vector;
generating an individual document vector for each document not opened during any retrieval session; and
combining the log-based document cluster vector and the individual document cluster vector.

5. The method recited in claim 3 wherein the second vector is used in place of a second document within the hybrid matrix wherein the second document forms a portion of the log-based document cluster.

6. The method recited in claim 3 wherein the step of clustering the documents using the hybrid matrix is performed using a content-based clustering technique.

7. The method recited in claim 4 wherein the step of clustering retrieval sessions into session clusters comprises the steps of:
generating a Boolean session vector for each retrieval session;
forming a matrix of the Boolean session vectors; and
applying a clustering algorithm to the matrix of the Boolean session vectors.

8. The method recited in claim 4 wherein the log-based document cluster is formed by concatenating the documents to be combined.

9. An apparatus for clustering documents, the apparatus comprising:
storage for storing retrieval session logs; and
processor, connected to the storage, programmed to
cluster the retrieval sessions into session clusters;
generate, for each session cluster, a log-based document cluster;

generate a log-based document cluster vector for each of the log-based document clusters;

generate an individual document vector for each document not opened during any retrieval session; and

cluster the documents using the log-based document cluster vectors and individual document vectors.

10. The apparatus recited in claim 9 wherein the documents are stored in the storage.

11. The apparatus recited in claim 9 further comprising:

memory, connected to the processor, for storage of a hybrid matrix comprising the log-based document cluster vectors and the individual document vectors.

12. A data processing system having session logs and documents, the system comprising:

a processor for executing program instructions; and

a media readable by the processor having a document clustering module having a plurality of instructions, which when executed by the processor, performs log-based clustering on the session logs to generate session clusters, converts the session clusters into a form suitable for content-based clusters, performs content-based clustering on the documents and session clusters in a form suitable for content-based clustering to generate document clusters with users' perspective.

13. The data processing system of claim 12 wherein the document clustering module further includes

a session vector generation module for receiving the session logs and based thereon for generating a session vector for each session log;

a session cluster generation module coupled to the session vector generation module for receiving the session vectors and based thereon for generating session clusters;

a hybrid matrix builder for receiving the documents, coupled to the session cluster generation module, for receiving the session clusters and based thereon for generating a hybrid matrix having at least one log-based document; and

a topic generation module coupled to the hybrid matrix builder for receiving the hybrid matrix and based thereon for generating document clusters with users' perspective.

14. The data processing system of claim 13 wherein the hybrid matrix builder further includes

a session document generation module for receiving session clusters and based thereon generates super documents; and

document modification module coupled to the session document generation module for receiving the super documents, for receiving the documents, and based thereon for generating the hybrid matrix.

15. The data processing system of claim 12 wherein the media is one of a floppy disk, compact disc, a volatile memory, and a non-volatile memory.

16. A machine readable memory device encoded with a data structure for clustering documents, the data structure having entries for a log-based document cluster vector generated from a log-based document cluster, and an individual document vector corresponding to a vector generated from a first document, the first document not belonging to any log based document cluster.